A LINEAR REGRESSION MODEL WITH BINOMIALLY DISTRIBUTED "ERRORS", WITH AN APPLICATION IN MARKETING

BU-200-M Thomas M. Beetle July, 1965

Abstract

Simple linear regression and associated normal distribution theory have been used to anlayze "brand preference" problems in marketing. The experimental procedure leads to a violation of the normal distribution and homogeneity of variance assumptions. This paper provides a model for the anlaysis of the experimental procedure. The simple linear regression approach is justified. Unbiased estimates of the variances of the least squares estimates, and normal approximation confidence interval estimates of the parameters are derived. Weighted least squares estimates are shown to be the maximum likelihood estimates. The optimum procedure for minimizing the variance of the regression coefficient estimate in a special case is discussed. In this case, the unweighted least squares estimate of the regression line is the minimum variance linear unbiased estimate.

Biometrics Unit, Plant Breeding Department, Cornell University

A LINEAR REGRESSION MODEL WITH BINOMIALLY DISTRIBUTED "ERRORS", WITH AN APPLICATION IN MARKETING

BU-200-M Thomas M. Beetle July, 1965

Introduction

Suppose that some product has a market of M people, and that $(1-\beta)p_1M$ of them prefer brand A over brand B and βM of them have no brand preference (where $0 \leq \beta \leq 1$ and $0 \leq p_1 \leq 1$). Consider the experiment of placing the two brands on a store shelf in such a way that brand A occupies a proportion p_2 of the space, and observing the proportion p of brand A sales ($0 \leq p_2 \leq 1$ and $0 \leq p \leq 1$). Assuming that p_2 is kept constant, that the patrons observed constitute a random sample from the population, and that "no preference" patrons select the product at random, then the expected value of the random variable p is

$$E(p) = (1-\beta)p_1 + \beta p_2 .$$

This expected value is derived by formulating the experimental situation as an urn model.

Since p_2 can be controlled, it seems reasonable to conduct an experiment as described above and to analyze the data as a linear regression of p on p_2 with intercept $(1-\beta)p_1$ and regression coefficient β.

Model: Let an urn contain M balls marked as follows:

$(1-\beta)p_1M$	red with the number 1
$(1-\beta)(1-p_1)M$	red with the number 0
βp_2M	black with the number 1
$\beta(1-p_2)M$	black with the number 0,

$0 \leq p_1 \leq 1$, $0 \leq p_2 \leq 1$, $0 \leq \beta \leq 1$.

Biometrics Unit, Plant Breeding Department, Cornell University

If a ball is drawn at random from the urn and the number on the ball, Y, is observed, then

$$P(Y=y) = \left\{ \left[\frac{(1-\beta)p_1M + (1-\beta)(1-p_1)M}{M} \cdot \frac{(1-\beta)p_1M}{(1-\beta)p_1M + (1-\beta)(1-p_1)M} \right] + \left[\frac{\beta p_2M + \beta(1-p_2)M}{M} \cdot \frac{\beta p_2M}{\beta p_2M + \beta(1-p_2)M} \right] \right\}^y$$

$$\left\{ \left[\frac{(1-\beta)p_1M + (1-\beta)(1-p_1)M}{M} \cdot \frac{(1-\beta)(1-p_1)M}{(1-\beta)p_1M + (1-\beta)(1-p_1)M} \right] + \left[\frac{\beta p_2M + \beta(1-p_2)M}{M} \cdot \frac{\beta(1-p_2)M}{\beta p_2M + \beta(1-p_2)M} \right] \right\}^{1-y}$$

$$= \left\{(1-\beta)p_1 + \beta p_2\right\}^y \left\{(1-\beta)(1-p_1) + \beta(1-p_2)\right\}^{1-y}$$

for y = 1 or 0.

If N balls are drawn with replacement and we let

$$p = \frac{1}{N} \sum_{i=1}^{N} Y_i \quad , \text{ then}$$

$$P(p=y) = \binom{N}{Ny} \left[(1-\beta)p_1 + \beta p_2\right]^{Ny} \left[(1-\beta)(1-p_1) + \beta(1-p_2)\right]^{N-Ny}$$

for Ny an integer and $0 \leq Ny \leq N$. This can be viewed as a binomial distribution with parameters N and $(1-\beta)p_1 + \beta p_2$.

As is well known, the expected value of p is

$$E(p) = (1-\beta)p_1 + \beta p_2 \ ,$$

and the variance of p is

$$V(p) = \frac{1}{N}\left[(1-\beta)p_1 + \beta p_2\right]\left[(1-\beta)(1-p_1) + \beta(1-p_2)\right] .$$

Changing notation to $p = Y$, $(1-\beta)p_1 = \alpha$ and $p_2 = X$, we can write the usual simple linear regression equation

$$Y = \alpha + \beta X + \epsilon \ ,$$

where ϵ is a binomial type random variable with mean 0 and variance $\frac{1}{N}\left[\alpha + \beta X\right]\left[1 - \alpha - \beta X\right]$.

Estimation: Let (Y_1, X_1, N_1), (Y_2, X_2, N_2), ..., (Y_M, X_M, N_M) denote M independent observations for the above regression model, where N_i is the number of observations used in computing Y_i; then there exists unique unbiased least squares estimates of α and β, providing that for some $i \neq j$, $X_i \neq X_j$. These estimates are the usual

$$(1) \quad \hat{\beta} = \frac{\sum_i (X_i - \bar{x})(Y_i - \bar{y})}{\sum_i (X_i - \bar{x})^2} \text{ , and}$$

$$(2) \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \text{ ,}$$

where $\bar{y}$ and $\bar{x}$ are the sample means.

In general, the Y_i random variables will have different variances. If we let

$$w_i = \frac{N_i}{(\alpha+\beta X_i)(1-\alpha-\beta X_i)} \text{ ,}$$

then, by the Gauss-Markoff Theorem[1], the minimum variance linear unbiased estimate for $\alpha + \beta X$ is $\tilde{\alpha} + \tilde{\beta}X$, where $\tilde{\alpha}$ and $\tilde{\beta}$ are the weighted least squares estimates:

$$(3) \quad \tilde{\beta} = \frac{\sum_i w_i \left(X_i - \frac{\sum_i w_i X_i}{\sum_i w_i}\right)\left(Y_i - \frac{\sum_i w_i Y_i}{\sum_i w_i}\right)}{\sum_i w_i \left(X_i - \frac{\sum_i w_i X_i}{\sum_i w_i}\right)^2}$$

$$(4) \quad \tilde{\alpha} = \frac{\sum_i w_i Y_i}{\sum_i w_i} - \tilde{\beta}\frac{\sum_i w_i X_i}{\sum_i w_i} \text{ .}$$

---

[1] Scheffe', Henry: The Analysis of Variance, John Wiley and Sons, Inc., New York, 1959, p. 14.

Let $\bar{y}_w = \dfrac{\sum_i w_i Y_i}{\sum_i w_i}$ and $\bar{x}_w = \dfrac{\sum_i w_i X_i}{\sum_i w_i}$ denote the weighted sample means.

Theorem 1. $\tilde{\alpha}$ and $\tilde{\beta}$ are maximum likelihood estimators.

Proof:
$$L = \prod_{i=1}^{M} \binom{N_i}{N_i Y_i} (\alpha + \beta X_i)^{N_i Y_i} (1 - \alpha - \beta X_i)^{N_i - N_i Y_i}$$

$$\log L = \sum_i \log \binom{N_i}{N_i Y_i} + \sum_i N_i Y_i \log (\alpha + \beta X_i) +$$

$$+ \sum_i (N_i - N_i Y_i) \log (1 - \alpha - \beta X_i)$$

$$\frac{\partial \log L}{\partial \alpha} = \sum_i \frac{N_i Y_i}{(\alpha + \beta X_i)} - \sum_i \frac{N_i - N_i Y_i}{(1 - \alpha - \beta X_i)}$$

$$\frac{\partial \log L}{\partial \beta} = \sum_i \frac{N_i Y_i X_i}{(\alpha + \beta X_i)} - \sum_i \frac{(N_i - N_i Y_i) X_i}{(1 - \alpha - \beta X_i)} .$$

Since $w_i = \dfrac{N_i}{(\alpha + \beta X_i)(1 - \alpha - \beta X_i)}$,

$$\frac{\partial \log L}{\partial \alpha} = \sum_i w_i Y_i (1 - \alpha - \beta X_i) - \sum_i w_i (1 - Y_i)(\alpha + \beta X_i)$$

$$\frac{\partial \log L}{\partial \beta} = \sum_i w_i Y_i X_i (1 - \alpha - \beta X_i) - \sum_i w_i (1 - Y_i) X_i (\alpha + \beta X_i).$$

Setting the partial derivatives equal to zero and solving for α and β results in $\alpha = \tilde{\alpha}$ and $\beta = \tilde{\beta}$. Q.E.D.

The variances of the unweighted and weighted estimates are

$$V(\hat{\beta}) = \frac{1}{\left[\sum_i (X_i - \bar{x})^2\right]^2} \sum_i \frac{(X_i - \bar{x})^2}{N_i} (\alpha + \beta X_i)(1 - \alpha - \beta X_i)$$

$$V(\hat{\alpha}) = \frac{1}{M^2} \sum_i \frac{1}{N_i} (\alpha + \beta X_i)(1 - \alpha - \beta X_i) + \bar{x}^2 V(\hat{\beta}) -$$

$$- \frac{2\bar{x}}{M \sum_i (X_i - \bar{x})^2} \sum_i \frac{(X_i - \bar{x})}{N_i} (\alpha + \beta X_i)(1 - \alpha - \beta X_i)$$

$$V(\tilde{\beta}) = \frac{1}{\sum_i w_i (X_i - \bar{x}_w)^2}$$

$$V(\tilde{\alpha}) = \frac{1}{\sum_i w_i} + \frac{1}{\sum_i w_i (X_i - \bar{x}_w)^2} \quad .$$

Since $E\left[\frac{Y_i - Y_i^2}{N_i - 1}\right] = \frac{1}{N_i - 1}\left[E(Y_i) - E(Y_i^2)\right] =$

$$= \frac{1}{N_i - 1}\left[(\alpha + \beta X_i) - \frac{(\alpha + \beta X_i)(1 - \alpha - \beta X_i)}{N_i} - (\alpha + \beta X_i)^2\right] =$$

$$= \frac{1}{N_i - 1}\left[\frac{N_i - 1}{N_i}(\alpha + \beta X_i)(1 - \alpha - \beta X_i)\right] = \frac{(\alpha + \beta X_i)(1 - \alpha - \beta X_i)}{N_i} ,$$

unbiased estimates of the variances of $\hat{\beta}$ and $\hat{\alpha}$ are

$$(5) \quad \widehat{V(\hat{\beta})} = \frac{1}{\left[\sum_i (X_i - \bar{x})^2\right]^2} \sum_i \frac{(X_i - \bar{x})^2}{N_i - 1}\left(Y_i - Y_i^2\right) , \text{ and}$$

$$(6) \quad \widehat{V(\hat{\alpha})} = \frac{1}{M^2} \sum_i \frac{1}{N_i - 1}\left(Y_i - Y_i^2\right) + \bar{x}^2 \widehat{V(\hat{\beta})} -$$

$$- \frac{2\bar{x}}{M\sum_i (X_i - \bar{x})^2} \sum_i \frac{(X_i - \bar{x})}{N_i - 1}\left(Y_i - Y_i^2\right) .$$

It can be shown that $\hat{\alpha}$ and $\hat{\beta}$ are both asymptotically normal with increasing N_i's. Hence, approximate $(1 - \delta)\%$ confidence interval estimates for $\hat{\alpha}$ and $\hat{\beta}$ are

$$(7) \quad CI(\hat{\alpha}) = \hat{\alpha} \pm \sqrt{\widehat{V(\hat{\alpha})}}\ z_\delta \text{ (normal), and}$$

$$(8) \quad CI(\hat{\beta}) = \hat{\beta} \pm \sqrt{\widehat{V(\hat{\beta})}}\ z_\delta \text{ (normal)} .$$

It is suggested that the N_i's be chosen in accordance with Cochran's

rules[2] when some hypothesized values for α and β are possible. E.g., if

$$H:(\alpha + \beta X_i) = .2, \text{ let } N_i \geq 200.$$

"Best" Conditions: The minimum variance linear unbiased estimate of $\alpha + \beta X$ is the weighted estimate $\tilde{\alpha} + \tilde{\beta}X$. However, in practice the weights w_i will not be known. In the special case where the estimates are computed on only two points, (Y_1, X_1, N_1) and (Y_2, X_2, N_2), the weighted estimates are equal to the unweighted estimates. That is,

$$\tilde{\beta} = \frac{w_1(X_1-\bar{x}_w)(Y_1-\bar{y}_w) + w_2(X_2-\bar{x}_w)(Y_2-\bar{y}_w)}{w_1(X_1-\bar{x}_w)^2 + w_2(X_2-\bar{x}_w)^2} =$$

$$= \frac{(X_2-X_1)(Y_2-Y_1)\, w_1w_2/(w_1+w_2)}{(X_2-X_1)^2\, w_1w_2/(w_1+w_2)} =$$

$$= \frac{(Y_2-Y_1)}{(X_2-X_1)} = \hat{\beta}\ , \quad \text{and}$$

$$\tilde{\alpha} = \frac{w_1Y_1 + w_2Y_2}{w_1 + w_2} - \frac{(Y_2-Y_1)}{(X_2-X_1)}\ \frac{w_1X_1 + w_2X_2}{w_1 + w_2} =$$

$$= \frac{1}{w_1 + w_2}\left[w_1(Y_1-\hat{\beta}X_1) + w_2(Y_2-\hat{\beta}X_2)\right] =$$

$$= \frac{w_1\hat{\alpha} + w_2\hat{\alpha}}{w_1 + w_2} = \hat{\alpha}\ .$$

Hence, in this case, the unweighted estimate $\hat{\alpha} + \hat{\beta}X$ is the minimum variance linear unbiased estimate. Also, if X_1' and X_2' are respectively the minimum and maximum values where linearity is expected to hold, then the following theorem shows that $V(\hat{\beta})$ is minimized by taking the observations (Y_1, X_1', N_1) and (Y_2, X_2', N_2). If α and β can be hypothesized, then N_1 and N_2 can be computed so as to minimize $V(\hat{\beta})$ for these two observations.

---

[2]Cochran, William G.: Sampling Techniques, John Wiley and Sons, Inc., New York, 1953, p. 41.

Theorem 2. Let X_1' and X_2' be the minimum and maximum values for which the linear regression model is valid. If (Y_1, X_1, N_1) and (Y_2, X_2, N_2) are to be observed with N_1 and N_2 fixed and $X_1' \le X_1 < X_2 \le X_2'$, then $V(\hat{\beta})$ is minimized by observing (Y_1, X_1', N_1) and (Y_2, X_2', N_2).

proof: $$V(\hat{\beta}) = \frac{1}{(X_1-X_2)^2}\left[\frac{(\alpha+\beta X_1)(1-\alpha-\beta X_1)}{N_1} + \frac{(\alpha+\beta X_2)(1-\alpha-\beta X_2)}{N_2}\right]$$

Let $u_1 = \alpha + \beta X_1$ and $u_2 = \alpha + \beta X_2$.

$$V(\hat{\beta}) = \frac{\beta^2}{N_1N_2(u_1-u_2)^2}\left[N_2u_1(1-u_1) + N_1u_2(1-u_2)\right]$$

Fix u_1 and let $u_2 \uparrow$. If $u_2 \ge \frac{1}{2}$, then $N_1u_2(1-u_2) \downarrow$ as $u_2 \uparrow$. Hence, in this range, $V(\hat{\beta}) \downarrow$.

Let $0 \le u_1 < u_2 < \frac{1}{2}$, and $u_2 = u_1 + \Delta$. Then,

$$V(\hat{\beta}) = \frac{\beta^2}{N_1N_2\Delta^2}\left[N_2u_1(1-u_1) + N_1(u_1+\Delta)(1-u_1-\Delta)\right]$$

$$= \frac{\beta^2}{N_1N_2}\left[\frac{-N_2u_1 - N_2u_1^2 + N_1u_1 - N_1u_1^2}{\Delta^2} + \frac{N_1(1-2u_1)}{\Delta} - N_1\right]$$

Since the numerators of the first two terms within the brackets are positive, $V(\hat{\beta}) \downarrow$ as $\Delta \uparrow$. Hence, if we fix X_1, $V(\hat{\beta})$ is minimized by letting $X_2 = X_2'$. By a similar argument we can show that if we fix X_2, $V(\hat{\beta})$ is minimized by letting $X_1 = X_1'$. Q.E.D.

Theorem 3. Suppose (Y_1, X_1, N_1) and (Y_2, X_2, N_2) are to be observed and $H:\alpha = a, \beta = b$ is posed. Let $N_1 + N_2 = N$, $C_i = (a + bX_i)(1 - a - bX_i)$ for $i = 1, 2$ and $C_1 < C_2$. Then $V(\hat{\beta})$ is minimized when

$$N_1 = \left[\frac{C_1\left(\sqrt{\frac{C_2}{C_1}} - 1\right)}{C_2 - C_1} N \right] \quad \text{or} \quad N_1 = \left[\frac{C_1\left(\sqrt{\frac{C_2}{C_1}} - 1\right)}{C_2 - C_1} N \right] + 1,$$

where $[\cdot]$ denotes the largest integer less than the number inside the brackets.

proof: $V(\hat{\beta}) = \frac{1}{(X_2-X_1)^2}\left(\frac{C_1}{N_1} + \frac{C_2}{N-N_1} \right)$. Let

$$f(N_1) = \frac{C_1}{N_1} + \frac{C_2}{N-N_1}$$

$$\frac{df}{dN_1} = \frac{N_1^2(C_2-C_1) + N_1(2NC_1) - N^2C_1}{N_1^2(N-N_1)^2} .$$

Set $\frac{df}{dN_1} = 0$ and solve the quadratic in N_1 numerator to obtain the positive root

$$N_1 = \frac{C_1\left(\sqrt{\frac{C_2}{C_1}} - 1\right)}{C_2 - C_1} N . \qquad \text{Q.E.D.}$$

When $C_2 = C_1$, let $N_1 = \left[\frac{N}{2} \right]$.

When (Y_1, X_1, N_1), (Y_2, X_2, N_2), ..., (Y_M, X_M, N_M) are to be observed, with $\Sigma N_i = N$, it is conjectured that $V(\hat{\beta})$ will be minimized by letting $M = 2$ and observing (Y_1, X_1', N_1) and (Y_2, X_2', N_2) for some $N_1 + N_2 = N$. This can be shown to be true for the proposed observations (Y_1, X_1', N_1), (Y_2, X_1', N_2), ..., (Y_K, X_1', N_K), (Y_{K+1}, X_2', N_{K+1}), ..., (Y_M, X_2', N_M).

Computing formulae: In the case where the linear regression is to be computed on the basis of two points, the following formulae will be convenient:

$$(9) \qquad \hat{\beta} = \frac{Y_2 - Y_1}{X_2 - X_1}$$

$$(10) \qquad \hat{\alpha} = \frac{Y_1X_2 - X_1Y_2}{X_2 - X_1}$$

$$(11) \qquad \widehat{V(\hat{\beta})} = \frac{(N_2-1)(Y_1-Y_1^2) + (N_1-1)(Y_2-Y_2^2)}{(N_1-1)(N_2-1)(X_2-X_1)^2}$$

$$(12) \qquad \widehat{V(\hat{\alpha})} = \frac{(Y_1-Y_1^2)}{(N_1-1)}\left[\frac{3X_1+X_2}{4(X_2-X_1)}\right] + \frac{(Y_2-Y_2^2)}{(N_2-1)}\left[\frac{X_1+3X_2}{4(X_1-X_2)}\right] + \bar{x}^2\,\widehat{V(\hat{\beta})}.$$

Conclusions: A linear regression model with binomially distributed "errors" has application in the "brand preference" problem. The experimental procedure outlined in the Introduction and the anlaysis derived from the model are appropriate for estimating in a population the proportion that has a brand preference, $1 - \beta$, and the proportion that prefer a given brand, $\alpha = (1-\beta)p_1$. The distribution of the estimate $\hat{p}_1 = \frac{\hat{\alpha}}{1-\hat{\beta}}$ has not been considered here.

The general least squares estimates are given in formulae (1) and (2) with unbiased estimates of their variances provided in formulae (5) and (6). Normal approximation confidence intervals are computed as in formulae (7) and (8). Cochran's rules for sample size are recommended.

In the special case where estimation is to be based on two observations, the estimates are maximum likelihood. Also, in this case, $\hat{\alpha} + \hat{\beta}X$ is the minimum variance linear unbiased estimate of $\alpha + \beta X$. The variance of $\hat{\beta}$ is minimized if the two values of the independent variable are chosen to be at the extremes of the range where the model is expected to hold. Under hypothesized values for α and β, the number of observations used in computing the proportions Y_1 and Y_2 can be determined so as to minimize $V(\hat{\beta})$ (see Theorem 3). Computing formulae for the two observation case are given in formulas (9) through (12).

Given that (Y_i, X_i, N_i) are to be observed with $\sum_{L=1}^{M} N_i = N$, it is conjectured that minimum variance for $\hat{\beta}$ will be achieved by letting $M = 2$, X_1 and X_2 the minimum and maximum of the range of linearity and $N_1 + N_2 = N$.